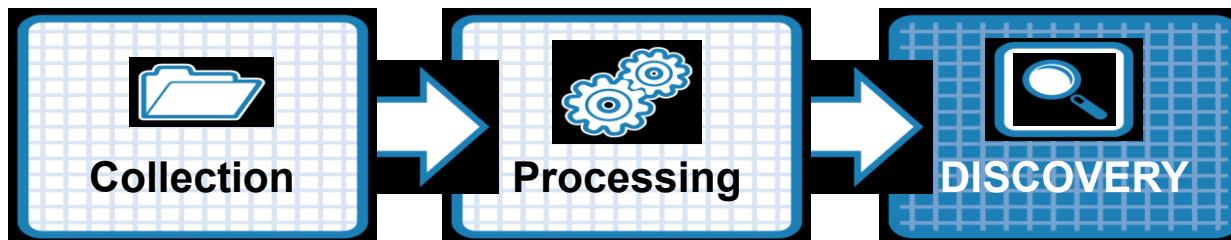




Data Discovery of Big and Diverse Observational Datasets – Options, Practices and Challenges

Presented To
GO-ESSP 2015 Workshop

Presented By
Giri Palanisamy
February 25, 2014



U.S. DEPARTMENT OF
ENERGY

Office of
Science

Presentation Summary

- Overview
- Data Architecture
- Data Discovery and Access
- PI Data Product Registration
- Data Citation & Linking Publications

The Atmospheric Radiation Measurement (ARM) mission is focused on global change

- Mission: The ARM Climate Research Facility, a DOE scientific user facility, provides the climate research community with strategically located in situ and remote sensing observatories designed to improve the understanding and representation, in climate and earth system models, of clouds and aerosols as well as their interactions and coupling with the Earth's surface.
- Vision: To provide a detailed and accurate description of the earth atmosphere in diverse climate regimes to resolve the uncertainties in climate and earth system models toward the development of sustainable solutions for the Nation's energy and environmental challenges.

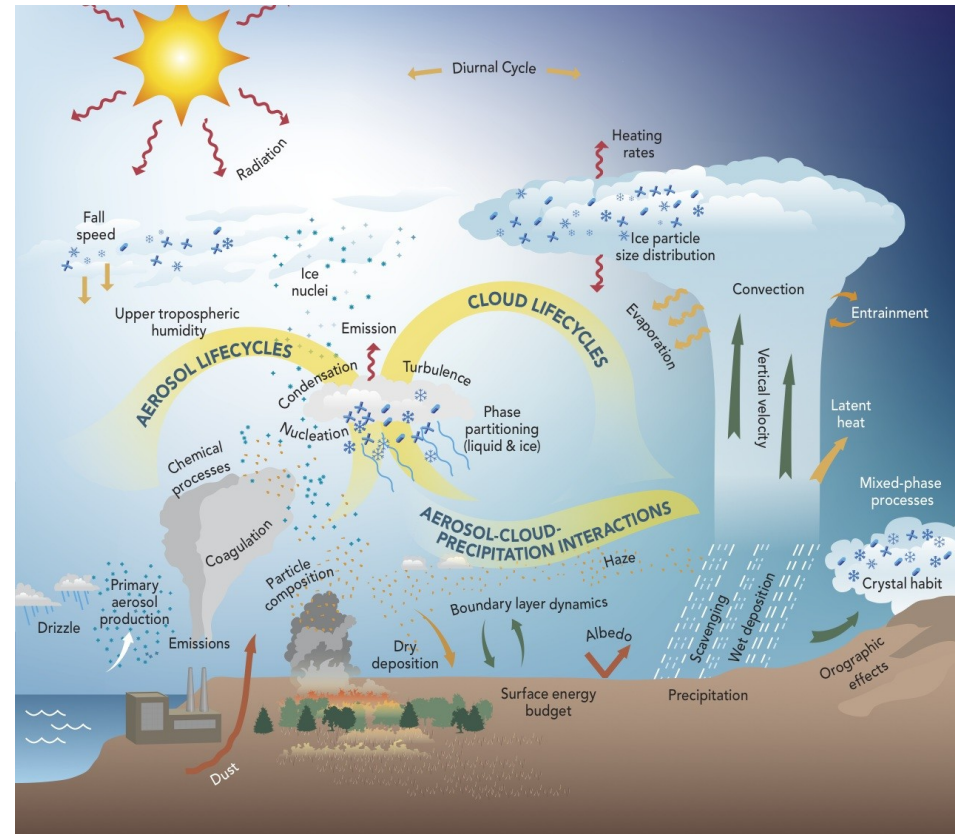
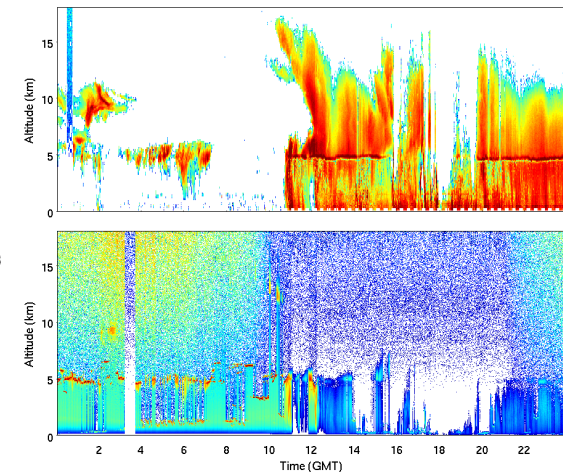
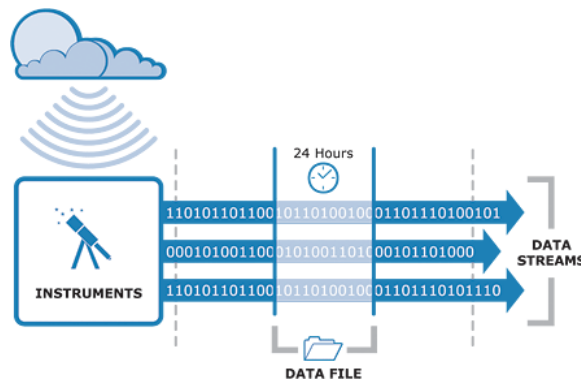
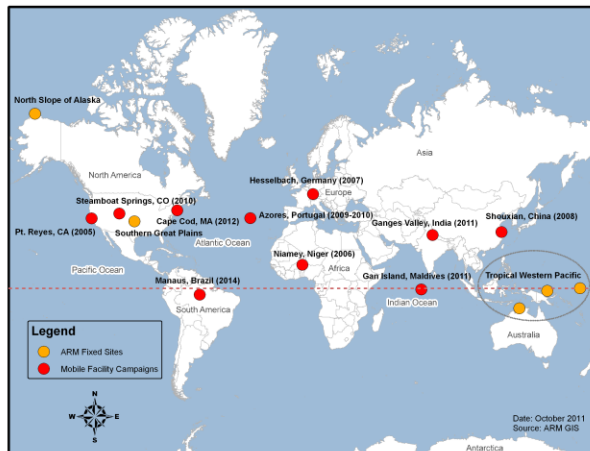


Image courtesy of the Atmospheric System Research program

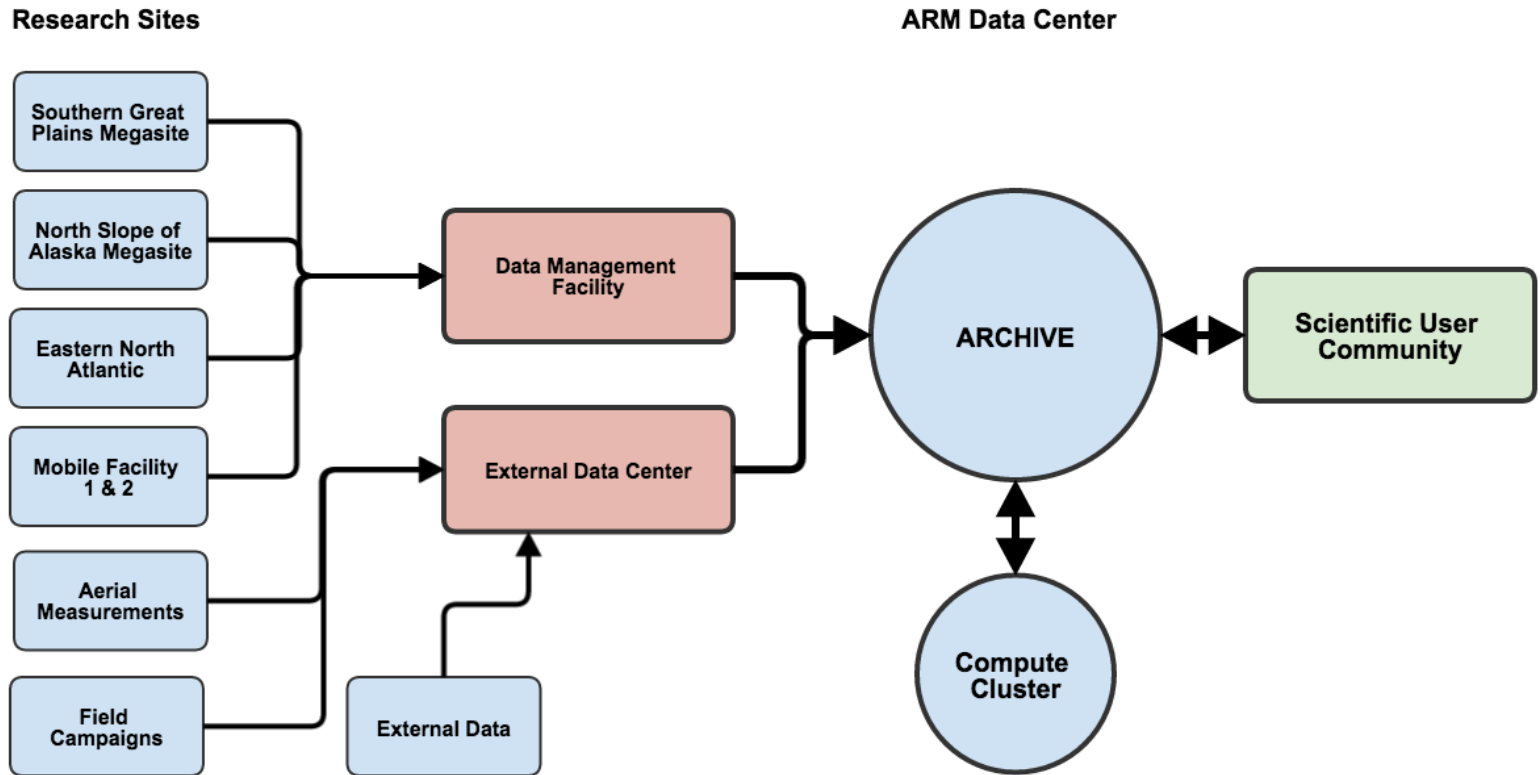
ARM is focused on providing high-quality data to support climate research

- Deploy measurement instrumentation
- Provide infrastructure to collect, process, preserve, and discover data
- Operate with excellence

Provide high-quality data to support global climate research

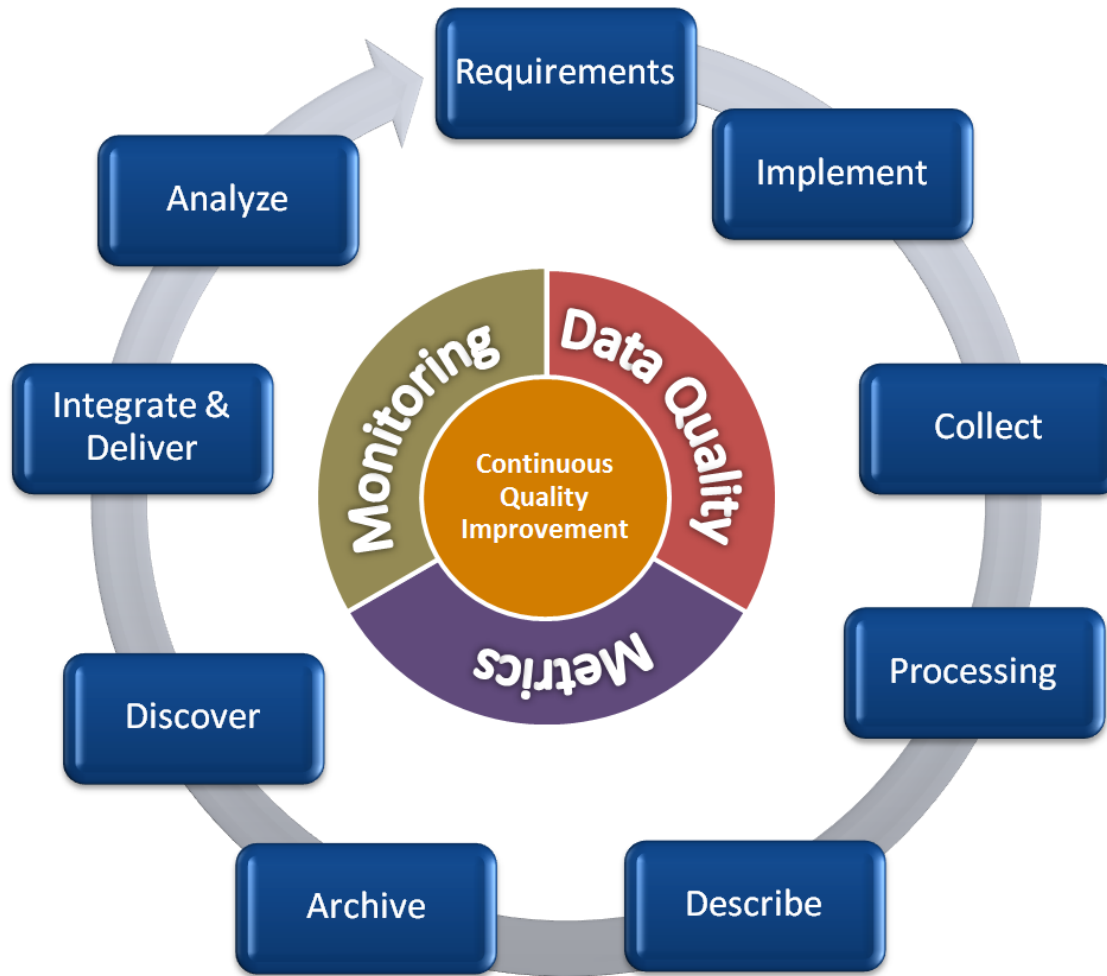


ARM Data Flow – The Big Picture



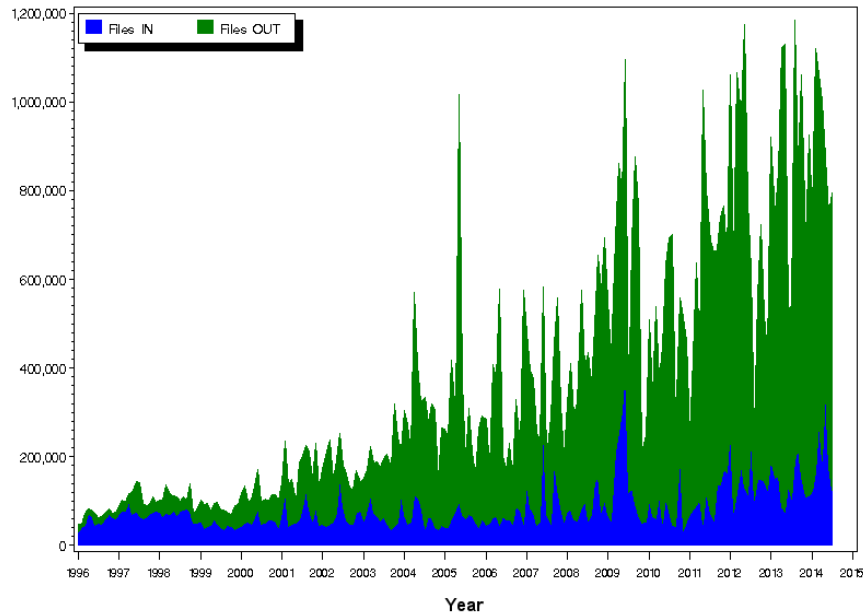
ARM Permanent Sites provide Long-Term Data.
Mobile Sites and Aircraft Increase Diversity.

ARM Data Lifecycle Architecture

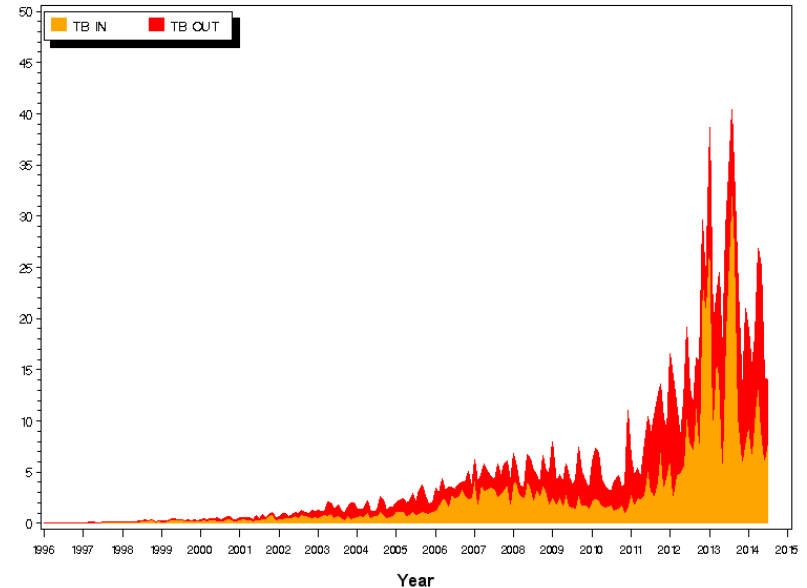


ARM Data Archival by the Numbers

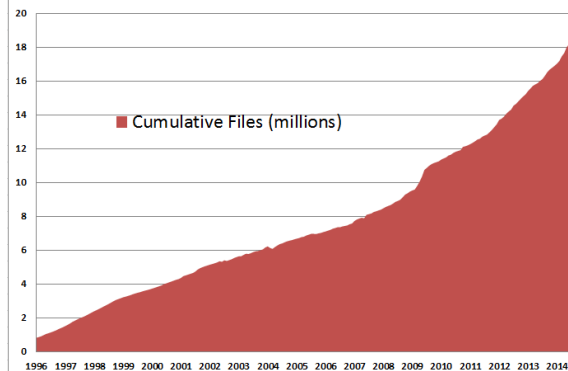
Number of Filenames In/Out per month



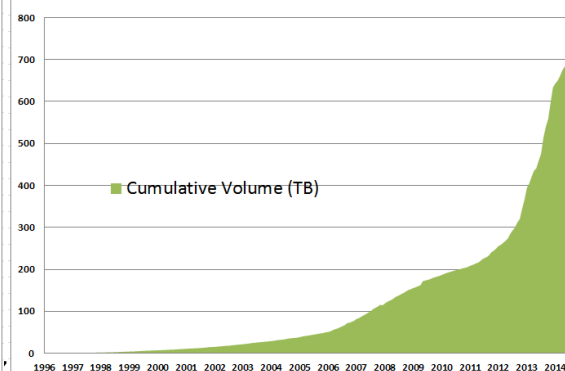
TB In/Out per month



History of Cumulative Number of Files Stored in ARM Archive



History of Cumulative Terabytes (TB) Stored in ARM Archive

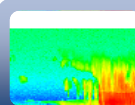
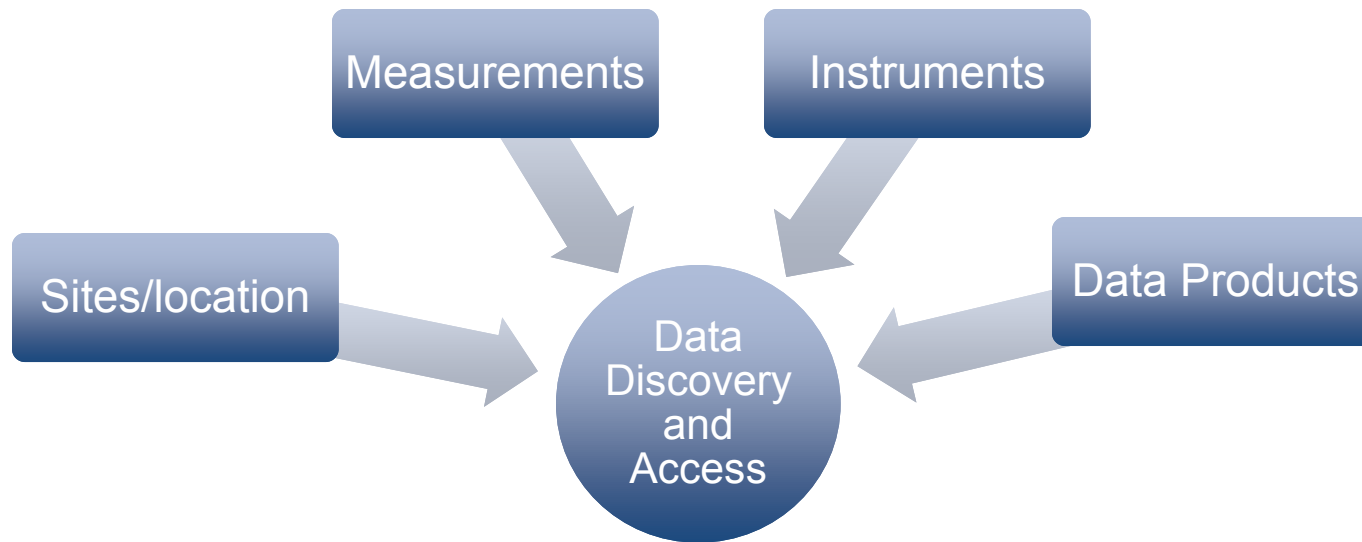
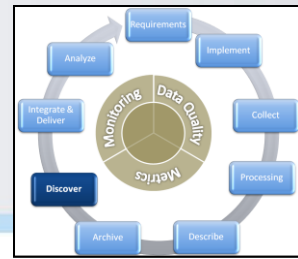


Data Discovery and Access

- ARM website is the primary place for:
 - Search
 - Discovery
 - Access
- Constantly upgraded to handle diverse ARM data

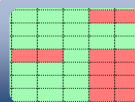
The screenshot displays the ARM Climate Research Facility website. At the top, the ARM logo is prominent, followed by the text 'CLIMATE RESEARCH FACILITY'. A navigation bar includes links for 'About', 'Science', 'Campaigns', 'Sites', 'Instruments', 'Measurements', 'Data', 'News', 'Publications', and 'Education'. A search bar is located in the top right corner. Below the navigation bar, a 'Recovery Act' section highlights ARM's efforts. A 'FEATURE' section dated 10.03.2011 titled 'AMIE, What You Wanna Do?' features a photograph of the ARM Mobile Facility and a text block describing the campaign. To the right, a 'USING OUR FACILITIES' section provides information on the preproposal cycle. Below this, a 'FIELD CAMPAIGNS' section lists ongoing projects: AMIE-GAN, GVAX, and MC3E. A 'FEATURED DATA' section on the left lists recent data releases with dates and brief descriptions. A 'News & Announcements' section on the right lists recent news items with dates and titles. The bottom of the page features the ARM logo and the U.S. Department of Energy Office of Science logo.

Multiple ways to search for ARM data



Data Plots

- Measurement plots
- Statistical plots



Data Quality

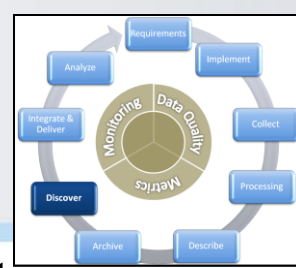
- DQ Report
- DQ Assessment



Data Citation

- DOIs for regular and PI data products
- Citation generation Tool

Data Discovery Tool



- Powerful data search capability to find and access ARM regular, PI and Field Campaign Data products
- Provides data availability in a timeline graphics
- Seamless access to data quality and data plots
- Provides options for data extraction and filtering based on data quality

ARM DATA DISCOVERY

CLIMATE RESEARCH FACILITY

aerosol optical depth

Search Results

To search for and request data, select a category, measurement, site, or source. Use the Start Date and End Date below to limit the data results timeline. Use the checkboxes below to add a data product to the Data Cart.

☐ ROUTINE DATA ☐ PI / CAMPAIGN DATA

2003-02-27

Showing 1-20 of 42 measurements

1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

AEROSOLBESTURN s1 @ SGP C1 // AEROSOL BEST ESTIMATE, FROM 1ST TURNER ALGORITHM

☐ - 1 Aerosol optical depth // Best estimate aerosol optical depth at 500 nm

☐ - 1 Aerosol optical depth // Best estimate aerosol optical depth at 355nm

AEROSOLBESTURN c1 @ SGP C1 // AEROSOL BEST ESTIMATE, FROM 1ST TURNER ALGORITHM

☐ - 1 Aerosol optical depth // Best estimate aerosol optical depth at 500 nm

☐ - 1 Aerosol optical depth // Best estimate aerosol optical depth at 355nm

AIPAVG1OQREN s1 @ MAO M1 // DERIVED: HOURLY AVERAGES OF AEROSOL INTENSIVE PROPERTIES FROM AOS, DELENE AND OGREN ET AL

☐ - Aerosol optical properties // Aerosol forcing per unit optical depth, 1 um size cut

☐ - Aerosol optical properties // Aerosol forcing per unit optical depth, 10 um size cut

AIP1OQREN s1 @ MAO M1 // DERIVED: AEROSOL INTENSIVE PROPERTIES FROM AOS, DELENE AND OGREN ET AL, 2001

☐ - Aerosol optical properties // Aerosol forcing per unit optical depth, 1 um size cut

☐ - Aerosol optical properties // Aerosol forcing per unit optical depth, 10 um size cut

MFRSRAOD1MICH s1 @ SGP E31 // MFRSR: DERIVED TOTAL & AEROSOL OPTICAL DEPTH FROM 1ST MICHALSKY ALGORITHM

☐ - Shortwave narrowband total downwelling irradiance // Radiation, shortwave, hemispheric irradiance, 870 nm wavelength

☐ - Shortwave narrowband diffuse downwelling irradiance // Radiation, shortwave, diffuse hemispheric irradiance, 615 nm wavelength

Consolidate data? ☐ No ☒ Yes

Primary source 30EBBR b1 @ SGP E32

File format NetCDF

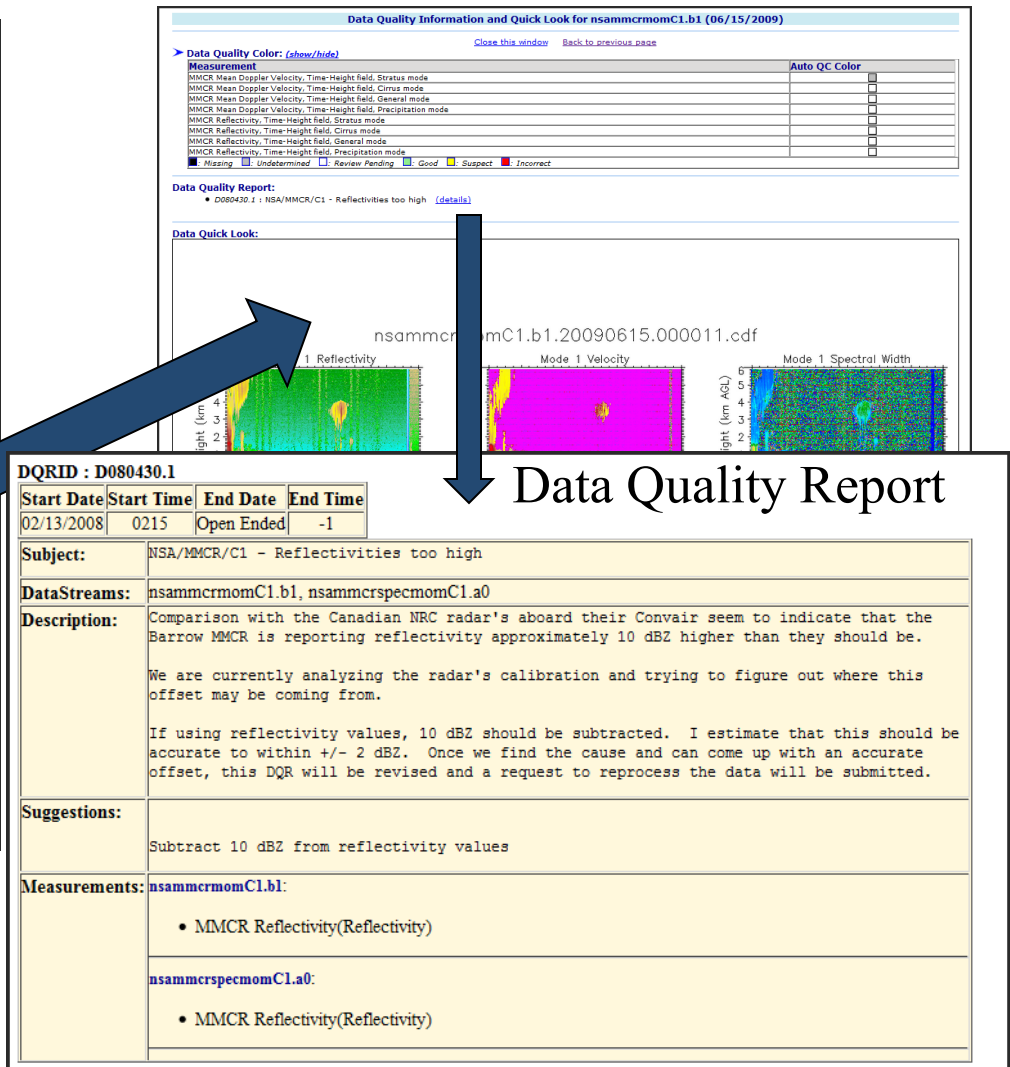
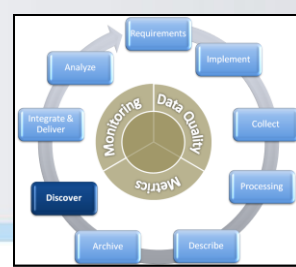
☐ to be ☒ Incorrect ☐ Suspect

Remove data points from DQR(s) known to be ☐ Incorrect ☐ Suspect

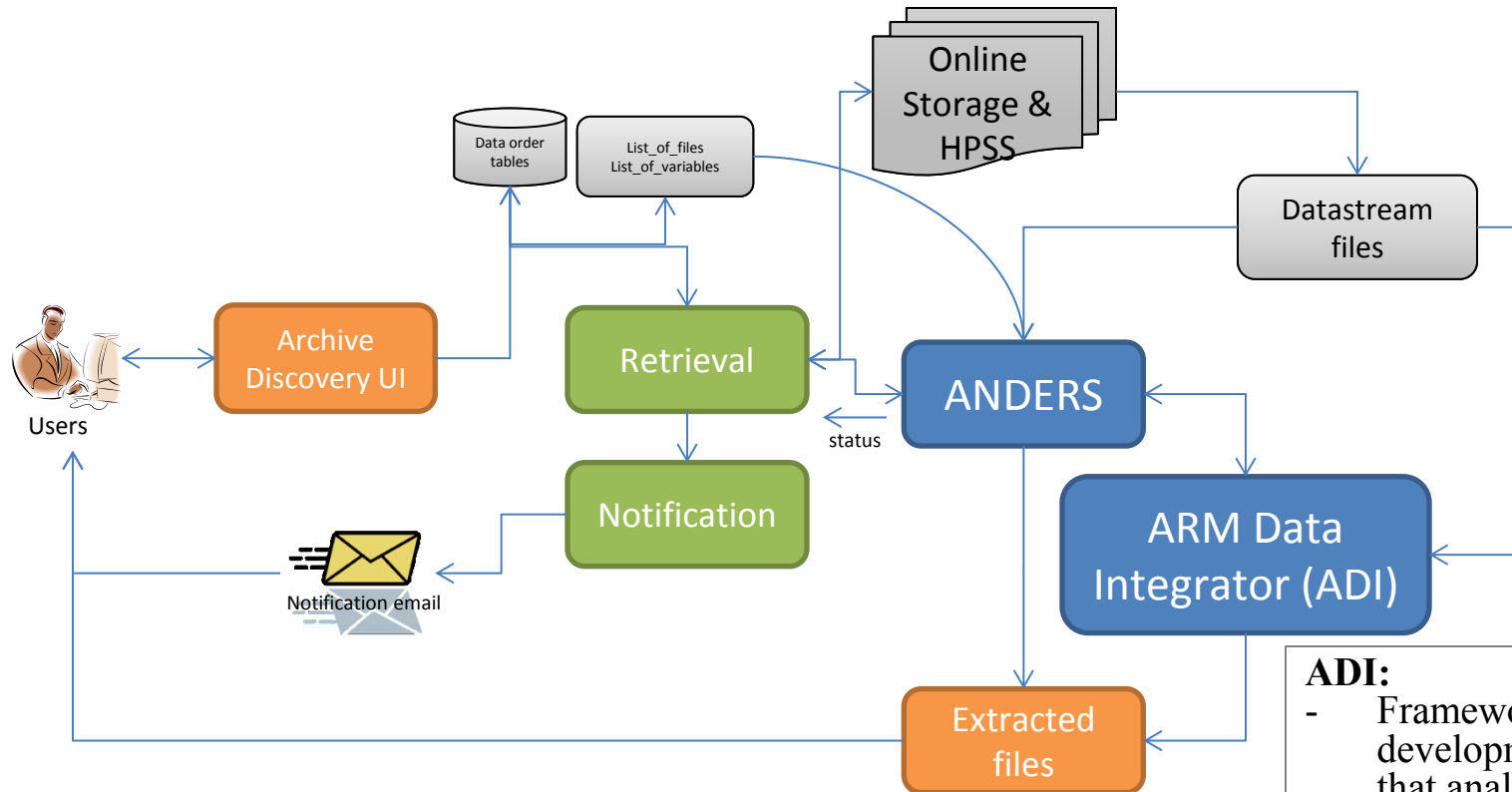
Office of Science

SUBMIT DATA REQUEST

Data Plots and Data Quality Reports – Helping Data Discovery



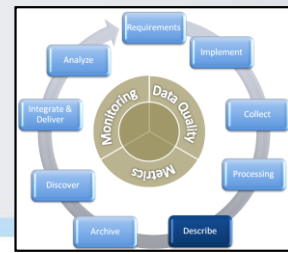
Archive Data Extraction Workflow



ADI:

- Framework to streamline the development of algorithms that analyze time-series data
- Suite of tools, libraries, data structures, and interfaces

ARM Data Product Registration and Submission Form (OME)



Data Quality

The Data Quality section of the metadata record is used to provide a general assessment of the quality of the dataset. There are four main components to this section:

Attribute Accuracy Report

An attribute is a defined characteristic of an entity within the dataset. E.g A data set might include the entity 'road' and have the attribute 'road type'.

How correct are the attribute values?

Attribute Accuracy refers to assessments as to how 'true' the attribute values may be - it may refer to field checks, cross-checks with other documents, statistical analysis values and parallel independent measures. It does not refer to the positional accuracy of the feature

Positional Accuracy Report

Consistency and Completeness Report

Logical Consistency Report provides an explanation for bad values or conditions and what tests and/or database QA/QC routines, if any, were used to check for data inconsistencies.

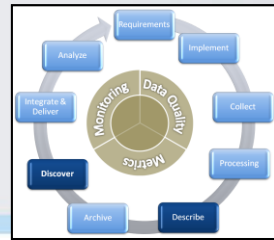
Does the dataset contain any bad values? If yes, what Quality Control/Quality Assurance (QA/QC) procedures were used?

E.g. do line intersect only where intended? Are polygon too small or lines too close?

Was there any factor affecting your research like cloud cover, precipitation e.t.c? Please explain:

- Data Type
- Description and keywords
- Contact information
- Data Quality
- When and Where
- Related Citations
- Analytical Tools
- Save, revisit and Submit

OME - Improving Data Discovery



After



Home | People | Site Index | Search arm.gov

Office of Science

About | Science | Campaigns | Sites | Instruments | Measurements | **Data** | News | Publications | Education

ARM.gov » Data » PI Data Products » CSSEF ARMBE

PI Product : CSSEF ARMBE

[RESEARCH DATA - EXTERNAL FUNDING]

The Climate Science for a Sustainable Energy Future (CSSEF) project is working to improve the representation of the hydrological cycle in global climate models, critical information necessary for decision-makers to respond appropriately to predictions of future climate. In order to accomplish this objective, CSSEF is building testbeds to implement uncertainty quantification (UQ) techniques to objectively calibrate and diagnose climate model parameterizations and predictions with respect to local, process-scale observations. In order to quantify the agreement between models and observations accurately, uncertainty estimates on these observations are needed. The DOE Atmospheric Radiation Measurement (ARM) program takes atmospheric and climate related measurements at three permanent locations worldwide. The ARM VAP called the ARM Best Estimate (ARMBE) [Xie et al., 2010] collects a subset of ARM observations, performs quality control checks, averages them to one hour temporal resolution, and puts them in a standard format for ease of use by climate modelers. ARMBE has been widely used by the climate modeling community as a summary product of many of the ARM observations. However, the ARMBE product does not include uncertainty estimates on the data values. Thus, to meet the objectives of the CSSEF project and enable better use of this data with UQ techniques, we created the CSSEFARMBE data set. For the current implementation of CSSEFARMBE, only a subset of the variables contained in ARMBE is included in CSSEFARMBE. CSSEFARMBE currently consists of only surface meteorological observations, though this may be expanded to include other variables in the future. The CSSEFARMBE VAP is focused on the ARM Southern Great Plains (SGP) site, and is produced for all extended facilities at SGP that contain surface meteorological equipment. This extension of the ARMBE data set to multiple facilities at SGP allows for better comparison between model grid boxes and the ARM point observations. In the future, CSSEFARMBE may also be created for other ARM sites. As each site has slightly different instrumentation, this will require additional development to understand the uncertainty characterization associated with instrumentation at those sites.

Comments?

We would love to hear from you!
Send us a note below or call us at
[1-888-ARM-DATA](tel:1-888-ARM-DATA).

Comments

SEND

Purpose

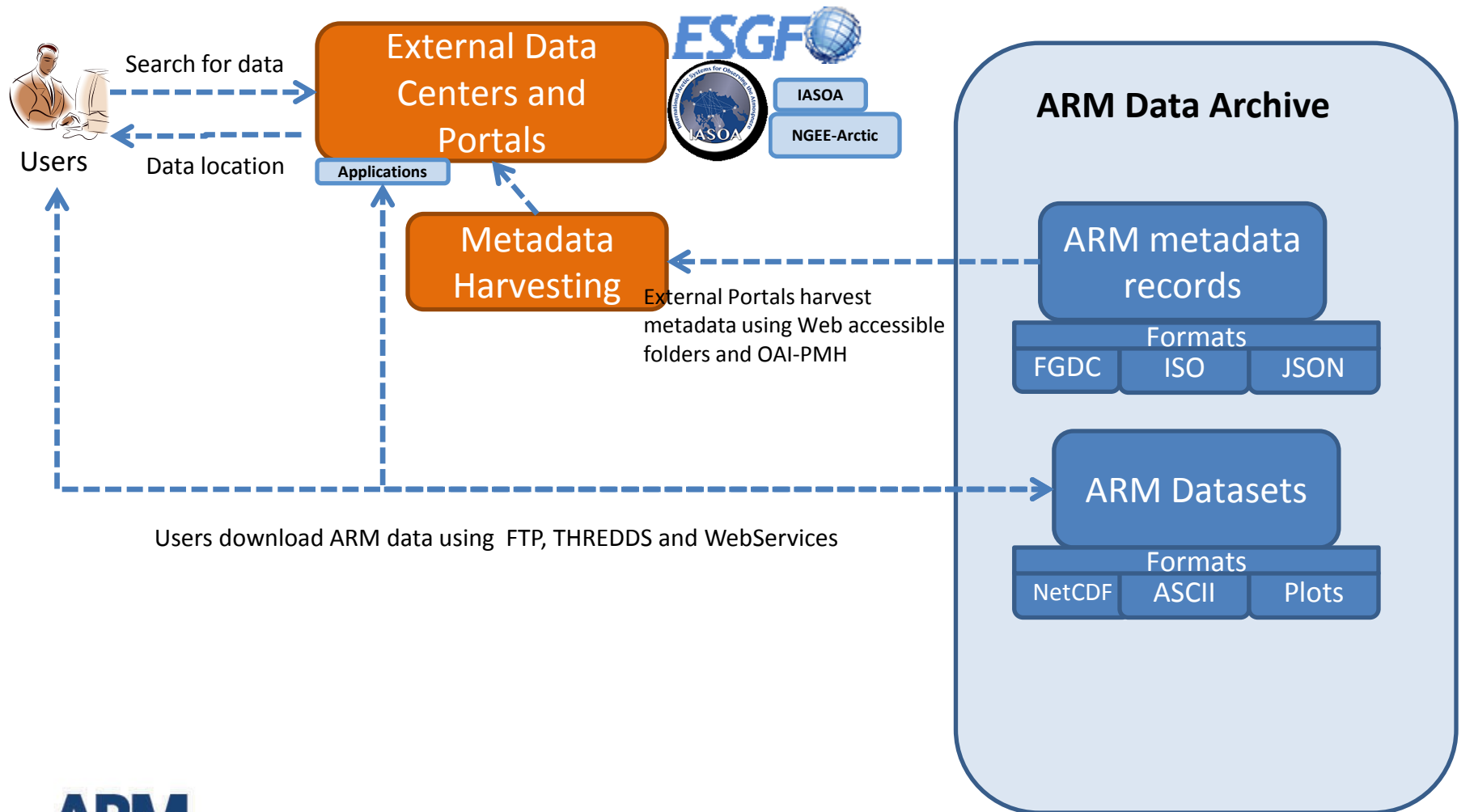
This data set was created for the Climate Science for a Sustainable Energy Future (CSSEF) model testbed project and is an extension of the hourly average ARMBE dataset to other extended facility sites and to include uncertainty estimates. Uncertainty estimates were needed in order to use uncertainty quantification (UQ) techniques with the data.

Data Details

DEVELOPED BY	Laura Riihimaki	
CONTACT	Laura Riihimaki laura.riihimaki@pnnl.gov (509) 375-6406 Richland, WA 99352	
	Krista Gaustad krista.gaustad@pnnl.gov (509) 375-5950 P.O. Box 999, K7-28 Richland, WA 99352	Sally McFarlane sally.mcfarlane@science.doe.gov (301) 903-0943 Climate and Environmental Sciences Division Washington, DC 20585
RESOURCE(S)	Data Directory	
DATA FORMAT	netcdf	
DATA USAGE	Positive and negative systematic, and random error components are given separately so that the uncertainties can be propagated appropriately when computing data averages. To propagate systematic uncertainties, a simple average can be used. Random errors should be propagated using the standard equation, square root[(random error) ² /Number of samples]. Error components should then be added in quadrature as described in the attached technical report.	
SITE INFORMATION	ARM SGP	
CONTENT TIME RANGE	2011.01.01 — 2011.12.31	

CONTENT TIME RANGE		2011.01.01 — 2011.12.31
SCIENTIFIC MEASUREMENTS	Measurement	Variables
	PRECIPITATION RATE	[expand]
	HORIZONTAL WIND	[expand]
	AIR TEMPERATURE	[expand]
	RELATIVE HUMIDITY	[expand]
	SURFACE AIR PRESSURE	[expand]
ATTRIBUTE ACCURACY	No formal attribute accuracy tests were conducted	
POSITIONAL ACCURACY	No formal positional accuracy tests were conducted	
DATA CONSISTENCY AND COMPLETENESS	Data set is considered complete for the information presented, as described in the abstract. Users are advised to read the rest of the metadata record carefully for additional details.	
FACTOR AFFECTING THE RESEARCH	Any data indicated bad by Data Quality Reports was removed from the data set.	
ACCESS RESTRICTION	No access constraints are associated with this data.	
USE RESTRICTION	No use constraints are associated with this data.	
FILE NAMING CONVENTION	(sss)cssefarmbe(FFF).c1.YYYYMMDD.HHMMSS.cdf where time indicates first time in file.	
DIRECTORY ORGANIZATION	each subfolder contains data from a different extended facility	
CITATIONS	Riihimaki LD, KL Gaustad, and SA McFarlane. 2012. Climate Science for a Sustainable Energy Future Atmospheric Radiation Measurement Best Estimate (CSSEFARMBE). PNNL-21831, Pacific Northwest National Laboratory, Richland, WA.	

ARM Metadata and Data Sharing With Other Portals



Users download ARM data using FTP, THREDDS and WebServices

ARM Data Citation Service

Goal

- Allow users to cite exact ARM data used in their research/publication
- Allow ARM to provide proper data citation credits to the PIs and collaborators
- Allows future data users and the project to easily track the data used in various articles

Solution:

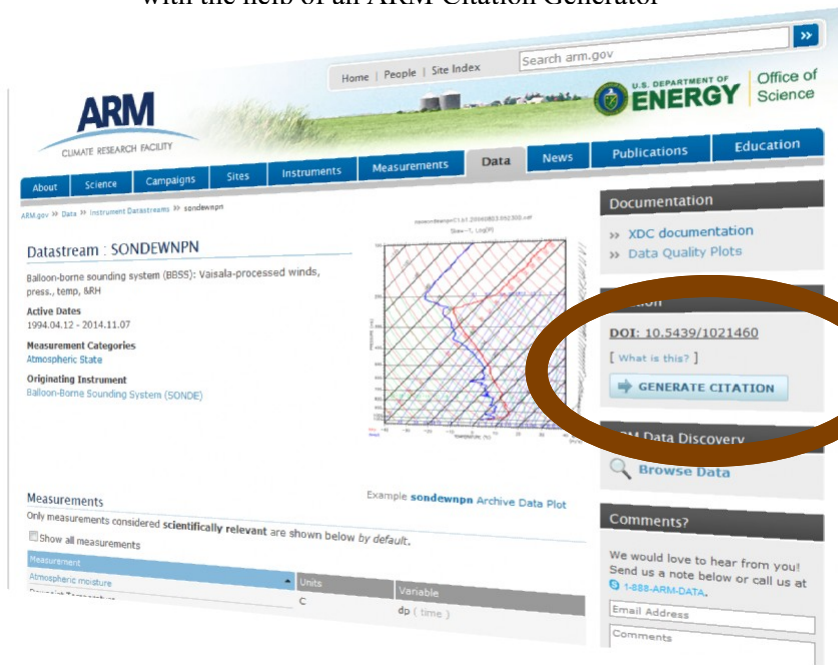
- DOIs are assigned at the data collection level
- A recommended citation allows users to cite the exact data with the help of an ARM Citation Generator

The Challenge

- Millions of data files from over 4000 data products
- These are continuous datastreams
- Large user community and complex use of data
- Data is also published via other portals

Example:

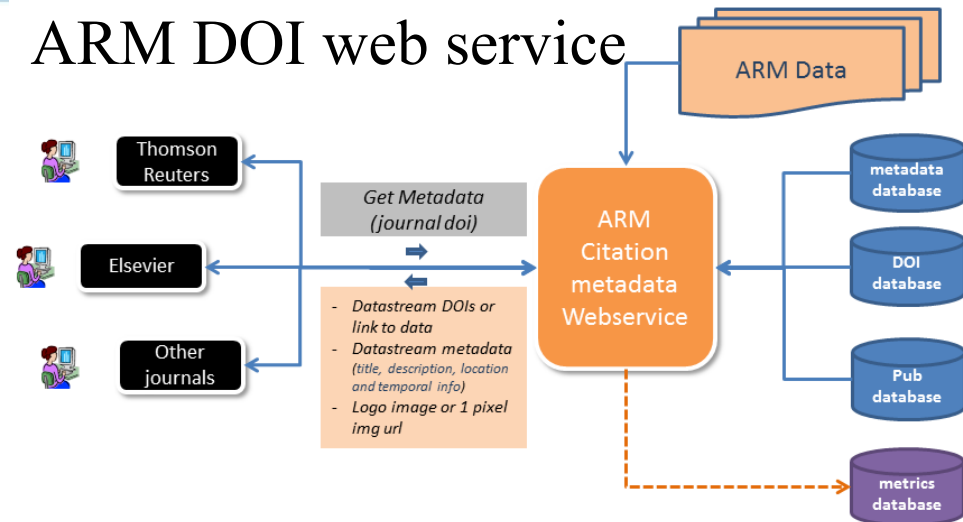
Atmospheric Radiation Measurement (ARM) Climate Research Facility. 1994, updated daily. SONDEWNPN. Oct. 2010–March 2011, 36° 36' 18.0" N, 97° 29' 6.0" W: Southern Great Plains Central Facility (C1). Compiled by R Coulter, J Prell, M Ritsche, and D Holdridge. ARM Data Archive: Oak Ridge, Tennessee, USA. Data set accessed 2011-04-13 at <http://dx.doi.org/10.5439/1021460>.



The screenshot shows the 'Generate Citation' form. The form includes fields for 'Author', 'Original Publication Date', 'Update Period', 'Location Accessed', 'Facility', 'Dates Used', 'Editor(s) or Compiler(s)', 'Date Accessed', and 'Citation(s)'. The 'Citation(s)' field contains the following text: 'Atmospheric Radiation Measurement (ARM) Climate Research Facility. 1994, updated hourly. Balloon-Borne Sounding System (SONDEWNPN). << start date used >> to << end date used >>, Summit Station, Greenland (SMT) << facility >>. Compiled by D. Holdridge, J. Kyrouac and R. Coulter. Atmospheric Radiation Measurement (ARM) Climate Research Facility Data Archive: Oak Ridge, Tennessee, USA. Data set accessed 2014-11-10 at <http://dx.doi.org/10.5439/1021460>'. The 'DONE' button is highlighted in red.

Discovering ARM Data from Publications

ARM DOI web service



The screenshot shows a ScienceDirect article page for "Remote Sensing of Environment". The article title is "Cloud model evaluation using radiometric measurements from the airborne multiangle imaging spectroradiometer (AirMISR)". The authors are Mikhail Ovchinnikov and Roger T. Marchand. The article is part of a special issue on "Multi-angle Imaging Spectroradiometer (MISR) Special Issue". The page includes an abstract, keywords, and a list of figures and tables. A red circle highlights the "Data For this Article" section, which contains the ARM logo and the text "Atmospheric Radiation Measurement Data".

The screenshot shows the ARM Data Selection Summary page. It displays a table of data streams, including "30EBBR b1 @ SGP E13" and "MFRSRAOD1MICH s1 @ PVC M1". The page includes a "Related Publications" section with a list of articles and a "Data Selection Summary" section with a table of data streams. The page also features a "Data Selection Summary" section with a table of data streams and a "Data Selection Summary" section with a table of data streams.

Thanks!



ARM Home Page: <http://www.arm.gov>

Key contacts:

DOE Program Manager: Sally McFarlane (Sally.McFarlane@science.doe.gov)

Technical Director: Jim Mather(Jim.Mather@pnnl.gov)

Chief Operating Officer: Jimmy Voyles (jimmy.voyles@pnnl.gov)

ARM Data Center: Giri Palanisamy (palanisamy@ornl.gov)